

국립국어원 추론_확신성 분석 말뭉치 2021

(버전 1.1)

- **자료명:** 국립국어원 추론_확신성 분석 말뭉치 2021
- **공개일**
 - (버전 1.0) 2022. 4. 1.
 - (버전 1.1) 2022. 8. 25.
 - 국립국어원 문어 말뭉치 추출 담화 삭제
- **자료 유형:** 텍스트
- **관련 사업:** 2021년 말뭉치 합의 분석 및 연구(2021)
- **자료 설명**
 - ※ 자세한 내용은 국립국어원 누리집 > 자료 > 연구·조사 자료 > ‘2021년 말뭉치 합의 분석 및 연구’ 사업 결과보고서 참조
- **내용**
 - 국립국어원 신문, 문어, 준구어 말뭉치에서 대상 담화 추출
 - 화자가 내포문에서 추출한 가설을 확인하는 정도에 대한 일반 언어 사용자의 직관 수집
 - 수집 결과 수치를 반영하여 ‘합의’, ‘중립’ 모순 표지 부착
 - SuperGLUE의 과제 중 CommitmentBank를 참고
 - ※ “추론 확신성 분석 말뭉치 2020”과의 비교
 - 다양한 유형(총 198개)의 보문소를 선정하여 담화를 추출
 - 추출 공정을 여러 단계로 하였으며, 작업자의 합의 절차를 거쳐 실험용 담화 결정
 - 대상 담화에 합의 취소 운용소는 포함되지 않을 수 있음
 - 설문조사 문항(가설)별 설문 조사 인원은 약 20명
 - 신뢰구간을 기준으로 ‘합의’, ‘모순’, ‘중립’ 표지를 부착
(단, 분류 표지보다는 함께 제공하는 실험 결과 수치 활용 권장)

· **분량:** 총 1562건 중 781건 공개*

	공개	비공개
규모	365	1,197

*이 말뭉치는 “2022년 국립국어원 인공지능 언어 능력 평가” 대회의 후보 과제로 검토중이므로 전체 분량의 50%만 공개

- 파일 형식: TSV(UTF-8 인코딩)
- 파일 수 및 크기: 파일 1개, 총 1.457KB

· 인용:

(국문) 국립국어원(202). 국립국어원 추론_확신성 분석 말뭉치 2021(버전 1.1). URL: <https://kli.korean.go.kr/corpus>

(영문) National Institute of Korean Language (2022). NIKL Korean Commitment Bank Corpus 2021 (v.1.1). URL: <https://kli.korean.go.kr/corpus>

· 데이터 구조와 예시

	설명	예시
idx	자료 번호	301019
source	담화 출처(파일명)	SBRW1900008528.1
genre	말뭉치 구분(문어, 신문, 구어, 대화)	구어
prev	선행 문맥	P1: 네. P5: 술을 많이 마셔도 회복이 될 겁니다. 옥희 선생님 같은 경우에도 P4: 네네 P5: 소시적에는 밤 새워서 이제 나이트에서 놀고 그러셔도 금방 회복이 되셨죠. P3: 아니 나 놀다니요. P3: 일하러 가죠. P5: 아~ 일을 하셨구나. P5: 예. P3: 예. P5: 우리가 이십대 삼십대 대체적으로 자기를 과신하고 잘못된 생활 습관 사십대 오십대 들어왔을 때 달라진 자기 몸의 환경에 맞게끔 자기 생활을 바꾸어야 되잖아요.
current	대상 문장	P5: 그렇게 바꾸어가면 만성 피로에 걸릴 일이 없거든요.
next	후행 문맥	P5: 그래서 저는 만성 피로에 걸리신 분들이 오면 반드시 이제 이거를 물어보는데요. P5: 만성 피로 뿌리 뽑는 법 땡땡땡을 높여라. P3: 면역력 P5: 면역력 P1,P2: 면역력 P5: 관계 있습니다. P1: 잠에 수면의 질을 높여라. P5: 처음부터 그렇게 정답을 맞추시면 P1: 잠깐만요. 정답이에요? all: {laughing}
pred	술어	없다
comp	보문소	르_일
ecomark	함의최소운용소	NA
eco	함의최소운용소 유형(부정, 의문, 조건, 양태, 없음)	없음
context+target	실험에서 응답자에게 제공된 담화(선행 문맥+대상 문장+후행 문맥)	우리가 이십대 삼십대 대체적으로 자기를 과신하고 잘못된 생활 습관 사십대 오십대 들어왔을 때 달라진 자기 몸의 환경에 맞게끔 자기 생활을 바꾸어야 되잖아요. 그렇게

		바꾸어가면 만성 피로에 걸릴 일이 없거든요. 그래서 저는 만성 피로에 걸리신 분들이 오면 반드시 이제 이것을 물어보는데요
question	실험에서 응답자에게 제공된 질문	다음 문장에 대해 위 보기의 필자(화자)는 어느 정도로 확신한다고 생각하십니까?
prompt	실험에서 응답자에게 제공된 내포문 기반 가설	사십대 오십대 들어왔을 때 달라진 자기 몸의 환경에 맞게끔 자기 생활을 바꾸어 가면 만성 피로에 걸린다
relation	담화 추출 작업자가 판단한 함의(확신성) 관계(함의, 중립, 모순)	모순
a1 ~ a21	응답자 응답 결과(a20, a21은 누락되었을 수 있음)	5, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 2, 1, 1, 1, 1, 1, 2
95% CI lower bound	신뢰구간 하위 95% 지점	0.96404
95% CI upper bound	신뢰구간 상위 95% 지점	1.83596
mean	실험결과치 평균	1.4
sd	실험결과치 표준편차	0.994722918
entailment	실험 결과치 기반 함의(확신성) 관계 95% 신뢰 구간 기준 5~7점:함의 95% 신뢰 구간 기준 3~5점:중립 95% 신뢰 구간 기준 1~3점:모순	모순
embagr	모문-내포문의 주어 일치 여부(y:일치, n:불일치)	n
embtns	내포문의 시제(현재, 과거, 미래)	미래
embtnsm	내포문의 시제 표지	없음
general_fact	총칭문 여부	y
embper	내포문 주어의 인칭(1,2,3, 알 수 없음)	3
mtrtns	모문 시제(현재, 과거, 미래)	현재
mtrtnsm	모문 시제 표지	없음
mtrper	모문 주어의 인칭(1,2,3, 알 수 없음)	3

수집·선별 담화, 국어학 분석 정보
실험 문항, 응답 결과, 결과값 정보

· 자료 내용 문의: 02-2669-9638